School of Economics and Management

STAN53, Statistics: High-dimensional Data Analysis, 7.5 credits

Statistik: Högdimensionell dataanalys, 7,5 högskolepoäng

Second Cycle / Avancerad nivå

Details of approval

The syllabus was approved by The Board of the Department of Statistics on 2022-08-29 (U 2022/536) and was last revised on 2024-03-07 (U 2024/138). The revised syllabus comes into effect 2024-03-08 and is valid from the autumn semester 2024.

General information

Second cycle level course in statistics. The course is recommended in a Master's degree in statistics. The course may also be taken as a single subject course or within other Master's programmes at Lund University.

Language of instruction: English

Main field of study	*Specialisation*
Statistics	A1N, Second cycle, has only first-cycle course/s as entry requirements

Learning outcomes

On a general level, the students should be able to understand the concept of analysing multivariate and high-dimensional data. They should be familiar with a basic minimum level of matrix competency, with general aspects of handling multivariate data, and understand the challenge in dealing with the data when their dimension is comparable with or bigger than the sample size. Upon successful completion of the course, the student will grasp the range of multivariate, dimension reduction, and regularisation techniques and will be able to summarise and interpret multivariate and high-throughput experimental data, apply the principal component analysis and factor analysis and demonstrate how these concepts are applied to data visualisation, will be able to use machine learning to high-throughput data, and draw appropriate conclusions.

Knowledge and understanding

For a passing grade, the student shall

- demonstrate knowledge of the range of multivariate techniques available with the focus on high-dimensional methods and high-throughput data,
- demonstrate an understanding of the challenges of using multivariate techniques for modern high-dimensional data, and
- demonstrate in-depth knowledge of regularisation methods, clustering analysis, and prediction algorithms such as k-nearest neighbours along with the concepts of training sets, test sets, error rates, and cross-validation.

Competence and skills

For a passing grade, the student shall

- be able to apply regularisation methods, clustering analysis, and prediction algorithms such as k-nearest neighbours along with the concepts of training sets, test sets, error rates, and cross-validation,
- be able to summarise results of analyses, including visualisation methods, and
- be able to explain the outcomes to a non-data scientist.

Judgement and approach

For a passing grade, the student shall

- identify proper techniques and computational techniques to perform statistical analysis of multivariate and high-dimensional empirical data.

Course content

The course starts with an introduction to matrices and multivariate normal distribution. It is followed by singular value decomposition and its geometric interpretation. Then the fundamentals of Principal Component Analysis including its functional formulation are covered. Prediction theory including prediction with high-dimensional predictors is presented next with emphasis on penalised regression and prediction. It is followed by a sparse singular value decomposition and linear discriminant analysis. The course concludes with large scale inference.

Course design

The course is made up of lectures, problem solving sessions, lab sessions, and a final seminar.

Assessment

The examination consists of quizzes, lab reports and a final project. The project is presented both in writing and in speech at a final seminar. At the seminar, the students will be questioned about their comprehension of the used methods.

The University views plagiarism very seriously, and will take disciplinary actions against students for any kind of attempted malpractice in examinations and assessments. Plagiarism is considered to be a very serious academic offence. The penalty that maybe imposed for this, and other unfair practice in examinations or assessments, i ncludes suspension from the University.

The examiner, in consultation with Disability Support Services, may deviate from the regular form of examination in order to provide a permanently disabled student with a form of examination equivalent to that of a student without a disability.

Grades

Grading scale includes the grades: U=Fail, E=Sufficient, D=Satisfactory, C=Good, B=Very Good, A=Excellent

A (Excellent) 85-100 points/percent. A distinguished result that is excellent with regard to theoretical depth, practical relevance, analytical ability and independent thought.

B (Very good) 75-84 points/percent. A very good result with regard to theoretical depth, practical relevance, analytical ability and independent thought.

C (Good) 65-74 points/percent. The result is of a good standard with regard to theoretical depth, practical relevance, analytical ability and independent thought and lives up to expectations.

D (Satisfactory) 55-64 points/percent. The result is of a satisfactory standard with regard to theoretical depth, practical relevance, analytical ability and independent thought.

E (Sufficient) 50-54 points/percent. The result satisfies the minimum requirements with regard to theoretical depth, practical relevance, analytical ability and independent thought, but not more.

F (Fail) 0-49 points/percent. The result does not meet the minimum requirements with regard to theoretical depth, practical relevance, analytical ability and independent thought.

To pass the course, the students must have been awarded the grade of E or higher.

The final grade is determined by the results of quizzes (40%), lab reports (40%) and the final project report (20%).

Entry requirements

90 credits in Statistics, or the equivalent.

Further information

This course replaces STAN41 Statistics: Multivariate Analysis. The two courses may not be combined in a degree.